Chapter 2
Basic Statistical Techniques

2.1 Basic statistical quantities: means and other moments

2.1.1 The Mean

The sample *mean* of a set of N values, x_i, where $i = 1, 2, 3...N$ is given by

$$\overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i. \tag{2.1}$$

The mean is the *first moment* about zero and should be distinguished from the *median*, which is the value in the center of the population (or the average of the two middle values if N is even).

The sample mean $\overline{x}$ is an *unbiased estimate* of the true population mean μ. An unbiased estimate implies that if we draw an infinite number of samples from the same underlying distribution, then the mean of all of the sample means will be equal to the underlying distribution's population mean μ.

In Practice.

- The median is a very useful quantity when the distribution of your dataset is not symmetric or contains outliers. For example, in **Fig. 2.1**, the median may be considered more representative of the data.

2.1.2 The Variance

The sample *variance* of a set of values, x_i, is given by

$$\overline{x'^2} = \frac{1}{N-1}\sum_{i=1}^{N} (x_i - \overline{x})^2 \tag{2.2}$$

where the prime denotes departures from the mean. The variance is the *second moment* about the mean. The division by $N - 1$ instead of the expected N is required for an unbiased estimate of the variance. An explanation for why this is can be found in any standard statistics textbook, but it basically boils down to the fact that the sample mean is itself an estimate and comes with its own uncertainties which gives the sample variance a low bias without the $N - 1$ correction.

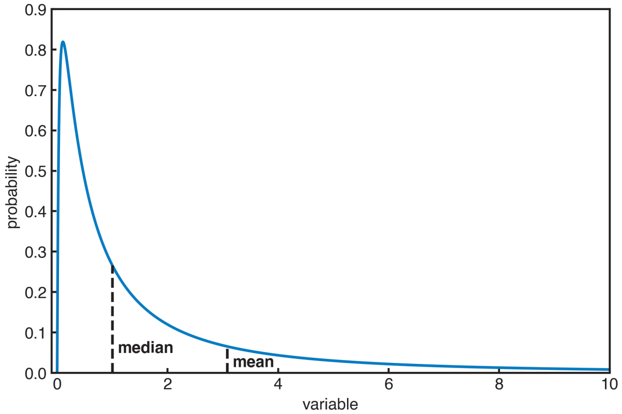

Figure 2.1 Comparison of the mean vs the median for a highly skewed distribution.

2.1.3 The Standard Deviation

The *standard deviation* is the square root of the variance and is often denoted as σ. The sample standard deviation of a set of values, x_i, is similarly defined as

$$s = \sqrt{\overline{x'^2}} = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})^2} \tag{2.3}$$

2.1.4 Higher Moments

We can define an arbitrary moment about the mean as

$$m_r = \frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})^r \tag{2.4}$$

so that m_2 is the variance, m_3 is the *skewness*, and m_4 is the *kurtosis*. Written in this way, note that m_2 is actually a biased estimate of the variance due to the division by N rather than $N - 1$.

These m_r moments can be standardized (non-dimensionalized) by defining

$$a_r = \frac{m_r}{\sigma^r} \tag{2.5}$$

where σ is the standard deviation. The first two standardized moments are zero and 1, but the third and fourth are the coefficients of skewness and kurtosis, which give information about the shape of the distribution.

The *coefficient of skewness*, a_3 indicates the degree of asymmetry of the distribution about the mean. If $a_3 > 0$ then the distribution is said to be *skewed to the right* and has a longer tail on the positive side. If $a_3 < 0$ then the distribution is said to be *skewed to the left* and has a longer tail on the negative side. **Fig. 2.2** shows examples of a positively and negatively skewed distribution.

The *coefficient of kurtosis* (Greek word for curved or arching), a_4, indicates the degree to which the distribution is spread about the mean value or the length of the tails. The kurtosis can be thought of as the "tailedness" of the distribution, and is typically compared with the kurtosis of the Normal distribution which has $a_4 = 3$. Thus, distributions with excess kurtosis ($a_4 > 3$) are very peaked about the mean with long tails and are called *leptokurtic* (Greek for *leptos*, meaning small or narrow) and distributions with $a_4 < 3$ are very flat about the mean with short tails and are called *platykurtic* (Greek *platys*, meaning broad or flat).

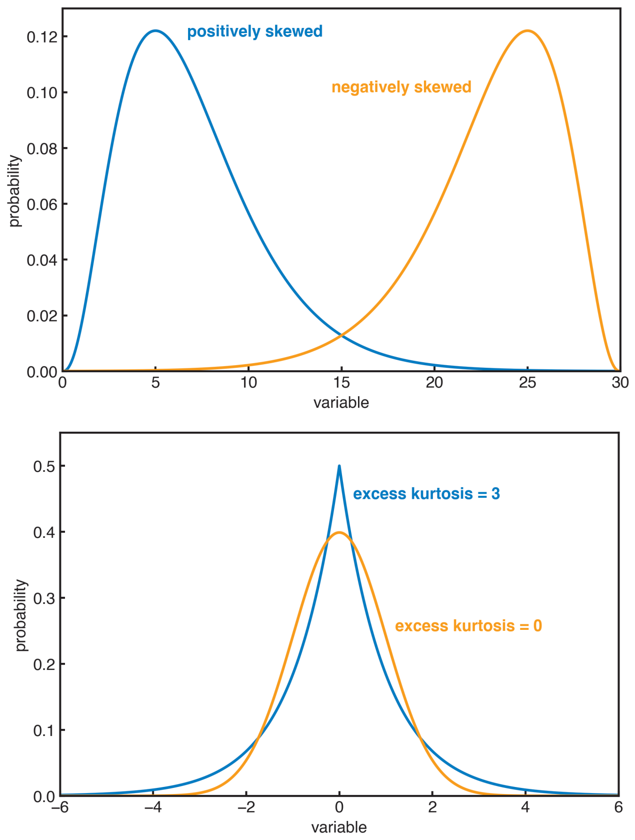

Figure 2.2 Examples of distributions with different skewness and kurtosis.

In Practice.

- In many software packages, the calculated kurtosis is actually the *excess kurtosis*, that is, the kurtosis minus 3 ($\alpha_4 - 3$) since 3 is the kurtosis of the Normal distribution. Thus, platykurtic distributions will have negative kurtosis, and leptokurtic positive kurtosis.

2.2 Probability Concepts and Theorems

2.2.1 Unions and Intersections of Probability - Venn Diagram

The probability of some event E happening is written as $\boldsymbol{Pr}(\mathsf{E})$. For example, E could be that you roll a die (a die is a cube with a different number, 1 through 6, on each side) and get a 2. If the die is fair, then

$$\boldsymbol{Pr}(\mathsf{E}) = \frac{1}{6}. \tag{2.6}$$

The probability of E not happening

$$\boldsymbol{Pr}\left(\widetilde{\mathsf{E}}\right) = 1 - \boldsymbol{Pr}(\mathsf{E}) \tag{2.7}$$

where $\widetilde{\mathsf{E}}$ is the event that E *does not* happen. In the case of rolling the fair die

$$\boldsymbol{Pr}\left(\widetilde{\mathsf{E}}\right) = 1 - \frac{1}{6} = \frac{5}{6}. \tag{2.8}$$

The probability that either or both of two events, E_1 and E_2, will occur is called the *union* of the two probabilities and is given by

$$\boldsymbol{Pr}(\mathsf{E}_1 \cup \mathsf{E}_2) = \boldsymbol{Pr}(\mathsf{E}_1) + \boldsymbol{Pr}(\mathsf{E}_2) - \boldsymbol{Pr}(\mathsf{E}_1 \cap \mathsf{E}_2) \tag{2.9}$$

where $\boldsymbol{Pr}(\mathsf{E}_1 \cap \mathsf{E}_2)$ is the probability that both events will occur, and is called the *intersection*. It is the overlap between the two probabilities and must be subtracted from the sum. This is easily seen via a Venn diagram in **Fig. 2.3**. The area inside the two event circles indicates the probability of the two events. The intersection between them gets counted twice when you add the two areas and so must be subtracted to calculate the union of the probabilities. If the two events are *mutually exclusive* (i.e. the circles do not overlap), then no intersection occurs.

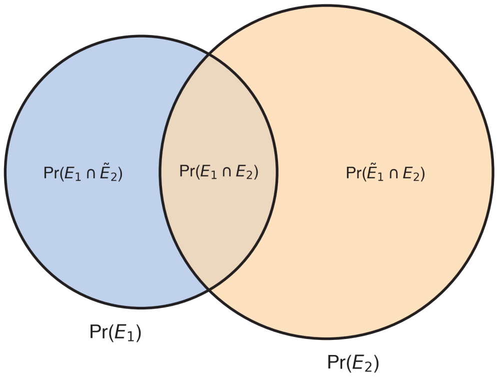
$\mathrm{Pr}(E_1 \cap \tilde{E}_2)$
$\mathrm{Pr}(E_1 \cap E_2)$
$\mathrm{Pr}(\tilde{E}_1 \cap E_2)$
$\mathrm{Pr}(E_1)$
$\mathrm{Pr}(E_2)$

Figure 2.3 Venn Diagram illustrating the intersection of two probabilities.

Another important concept is *conditional probability*. We write the probability that E_2 will occur given that E_1 has occurred as

$$\boldsymbol{Pr}(\mathsf{E}_2|\mathsf{E}_1) = \frac{\boldsymbol{Pr}(\mathsf{E}_1 \cap \mathsf{E}_2)}{\boldsymbol{Pr}(\mathsf{E}_1)} \tag{2.10}$$

Moving terms around, one can also obtain a formula for the probability that both events will occur, and this is called the multiplicative law of probability,

$$\boldsymbol{Pr}(\mathsf{E}_1 \cap \mathsf{E}_2) = \boldsymbol{Pr}(\mathsf{E}_2|\mathsf{E}_1)\,\boldsymbol{Pr}(\mathsf{E}_1) = \boldsymbol{Pr}(\mathsf{E}_1|\mathsf{E}_2)\,\boldsymbol{Pr}(\mathsf{E}_2)\,. \tag{2.11}$$

If E_1 and E_2 are *independent events*, that is, their probabilities do not depend on one another, then

$$\boldsymbol{Pr}(\mathsf{E}_1|\mathsf{E}_2) = \boldsymbol{Pr}(\mathsf{E}_1) \tag{2.12}$$

$$\boldsymbol{Pr}(\mathsf{E}_2|\mathsf{E}_1) = \boldsymbol{Pr}(\mathsf{E}_2) \tag{2.13}$$

and so

$$\boldsymbol{Pr}(\mathsf{E}_1 \cap \mathsf{E}_2) = \boldsymbol{Pr}(\mathsf{E}_1)\,\boldsymbol{Pr}(\mathsf{E}_2) \tag{2.14}$$

(2.14) is the definition of statistical *independence*.

Worked Example 2.1.
If the probability of getting heads on a coin flip is 0.5, and one coin flip is independent of every other coin flip, then, using (2.14), the probability of getting N heads in a row is 0.5^{N}.

Worked Example 2.2.
The probability of it raining on Monday is 60%. But, you know from looking at historical records that the probability of it raining the day after it rains is 80% (it is more likely than not to rain the day after it rains). So, whether it rains on Tuesday is dependent on whether it rains on Monday. What is the probability it will rain Monday and Tuesday?
....

$$\mathsf{M} = \text{event that it rains Monday} \tag{2.15}$$

$$\mathsf{T} = \text{event that it rains Tuesday} \tag{2.16}$$

$$\boldsymbol{Pr}(\mathsf{M} \cap \mathsf{T}) = \boldsymbol{Pr}(\mathsf{T}|\mathsf{M}) \cdot \boldsymbol{Pr}(\mathsf{M}) = 0.8 \cdot 0.6 = 48\% \tag{2.17}$$

2.2.2 Bayes Theorem

Theorem 2.1 (Bayes Theorem). *Let $\mathsf{E}_i, i = 1, 2, 3...\mathsf{N}$ be a set of N events, each with positive probability, such that E includes all possibilities in a set S and the events are mutually exclusive. Then, for any event B defined on S, with $\boldsymbol{Pr}(\mathsf{B}) > 0$,*

$$\boldsymbol{Pr}(\mathsf{E}_j|\mathsf{B}) = \frac{\boldsymbol{Pr}(\mathsf{B}|\mathsf{E}_j)\,\boldsymbol{Pr}(\mathsf{E}_j)}{\sum_{i=1}^{N}\boldsymbol{Pr}(\mathsf{B}|\mathsf{E}_i)\,\boldsymbol{Pr}(\mathsf{E}_i)} \tag{2.18}$$

Bayes Theorem may at first appear quite complicated, but in fact, we have already discussed all of the pieces that go into its derivation. We start with the conditional probability of an event E given that an event B has occurred:

$$\boldsymbol{Pr}(\mathsf{E}|\mathsf{B}) = \frac{\boldsymbol{Pr}(\mathsf{E} \cap \mathsf{B})}{\boldsymbol{Pr}(\mathsf{B})}. \tag{2.19}$$

This can be rearranged as

$$\boldsymbol{Pr}(\mathsf{E} \cap \mathsf{B}) = \boldsymbol{Pr}(\mathsf{E}|\mathsf{B})\,\boldsymbol{Pr}(\mathsf{B})\,. \tag{2.20}$$

If the E_i cover all possible outcomes, with a little thought one can see that the following must be true:

$$\boldsymbol{Pr}(\mathsf{B}) = \sum_{i=1}^{N}\boldsymbol{Pr}(\mathsf{B}|\mathsf{E}_i)\,\boldsymbol{Pr}(\mathsf{E}_i)\,. \tag{2.21}$$

Plugging (2.21) into the denominator of (2.19) gives us (2.18).

In Practice.

- In general, Bayes Theorem takes information about the $\boldsymbol{Pr}(\mathsf{A}|\mathsf{B})$ and turns it into information about the $\boldsymbol{Pr}(\mathsf{B}|\mathsf{A})$.

Worked Example 2.3.

You recently started measuring daily precipitation in Argentina to study extreme precipitation events in the area. Past experience at the site indicates that 5% of the days exhibit what you consider dangerous amounts of precipitation (e.g. lead to landslides, crop damage, etc.).

You are testing a new rain gauge that measures daily precipitation totals and then logs it into a computer. Unfortunately, the particular gauge in question has some reliability problems. Your gauge indicates extreme precipitation on only 95% of the days that extreme downpours actually occur. Furthermore, your gauge also incorrectly indicates extreme precipitation on 10% of the days when the actual precipitation was below what you consider extreme.

What is the probability that a day for which the gauge indicated extreme precipitation did not have extreme precipitation?

....

If we let E denote the event of extreme precipitation, and M denote the event where the gauge flags extreme precipitation, then we want to know $\boldsymbol{Pr}\left(\widetilde{\mathsf{E}}|\mathsf{M}\right)$. In this case, Bayes Theorem takes the form of

$$\boldsymbol{Pr}\left(\widetilde{\mathsf{E}}|\mathsf{M}\right) = \frac{\boldsymbol{Pr}\left(\mathsf{M}|\widetilde{\mathsf{E}}\right)\,\boldsymbol{Pr}\left(\widetilde{\mathsf{E}}\right)}{\boldsymbol{Pr}\left(\mathsf{M}|\widetilde{\mathsf{E}}\right)\,\boldsymbol{Pr}\left(\widetilde{\mathsf{E}}\right) + \boldsymbol{Pr}(\mathsf{M}|\mathsf{E})\,\boldsymbol{Pr}(\mathsf{E})} \tag{2.22}$$

$$= \frac{0.1 \cdot 0.95}{0.1 \cdot 0.95 + 0.95 \cdot 0.05} \approx 0.67 \tag{2.23}$$

Thus, a Bayesian would conclude that there is a 67% chance that the gauge is wrong and that extreme precipitation did not occur.

2.2.3 Probability philosophies: frequentist vs Bayesian views

While there are a wide range of philosophies on the meaning of probability, two general philosophies are discussed most frequently: the frequentist viewpoint, and the Bayesian viewpoint.

A *frequentist* approach takes the following form: If you have some large number of opportunities for an event to occur, then the number of times that event actually occurs, divided by the number of opportunities for it to occur is the probability. The probability varies between zero and one. The frequentist view has a solid foundation in the *Weak Law of Large Numbers* which states that if you have an event E that occurs N_E times in N trials, then $\mathsf{N}_\mathsf{E}/\mathsf{N}$ converges to the probability of event E occurring as the number of trials goes to infinity.

An alternative philosophy is attributed to Rev. Thomas Bayes (1701-1761), who figured that in many cases one is unlikely to have a large enough sample with which to measure the frequency of occurrence, and so, one must take a more liberal view. *Bayesian* inference is given that name for its frequent use of Bayes Theorem, which it uses to take into account *a priori* information, that may not be derivable from a frequentist point of view.

While the frequentist viewpoint is often found in the scientific literature in association with hypothesis testing and p-values, many recent articles have come out arguing against this approach due its broad misuse and the prevalence of "p-hacking" (e.g. Nuzzo, 2014; Goodman, 2001). Both the frequentist and the Bayesian approaches can be valid and useful, if done carefully and objectively. Bayesian analysis can be useful if you only have a small sample and you have prior information that you feel is reliable. New data can then be added to improve the estimate of probabilities. A weakness might be that this prior information could be subjective, and the methods of Bayesian analysis are a bit more complex. The Frequentist approach is simple to apply and works well if a large amount of data is available. Which approach to choose may depend on the problem at hand. In all cases one must be alert to the possibilities of errors in the logic or application of statistical tests.

Worked Example 2.4.
Sometimes, the frequentist approach and the Bayesian approach can result in different conclusions, as demonstrated by returning to our previous example of the faulty rain gauge.

....

Frequentist Approach: A frequentist would conclude that the probability that extreme precipitation did not occur is 10%, since this is the probability that the gauge incorrectly flags extreme precipitation when none actually occurred.

Bayesian Approach: The Bayesian approach would take into account the background rate of extreme precipitation and plug everything into Bayes Theorem. Taking this Bayesian approach, we previously concluded that there is a 67% chance that the gauge is wrong and that extreme precipitation did not occur.

The reason the two approaches result in such wildly different answers is that the Bayesian approach took into account information that the frequentist approach did not. Namely, the frequency with which extreme precipitation actually occurs.

2.3 Probability Distributions

The probability that a randomly selected value of a random variable x falls between the limits a and b is

$$\boldsymbol{Pr}(a \leqslant x \leqslant b) = \int\limits_a^b f(x)dx \tag{2.24}$$

This expression defines the *probability density function (PDF)*, $f(x)$, in the continuous case. Note that the probability that x is exactly equal to some value c is exactly zero.

To be a probability density function, $f(x)$ must satisfy the following criteria:

$$\int\limits_{-\infty}^{\infty} f(x)dx = 1 \tag{2.25}$$

$$f(x) \geqslant 0 \text{ for all } x \tag{2.26}$$

The moments about the mean of the distribution can be obtained directly from the probability density function using the following formula,

$$m_r = \int\limits_{-\infty}^{\infty} (x-\mu)^r f(x)dx, \tag{2.27}$$

where μ is the true, population mean.

The *cumulative distribution function (CDF)*, $F(x)$, is defined as the probability that a random variable assumes a value less than x,

$$F(x) = \int\limits_{-\infty}^{x} f(t)dt. \tag{2.28}$$

The probability density function and the cumulative density function are linked via the fundamental theorem of calculus and it is straightforward to show that

$$\frac{dF}{dx} = f(x), \tag{2.29}$$

and

$$\boldsymbol{Pr}(a \leqslant x \leqslant b) = \int_a^b f(x)dx = F(b) - F(a). \tag{2.30}$$

In Practice.

- The probability density function and cumulative density function of a finite data set can be approximated by the smoothed histogram of the data. One common method for smoothing is called the *kernel density estimation*, an example of which is given in **Fig. 2.4**.

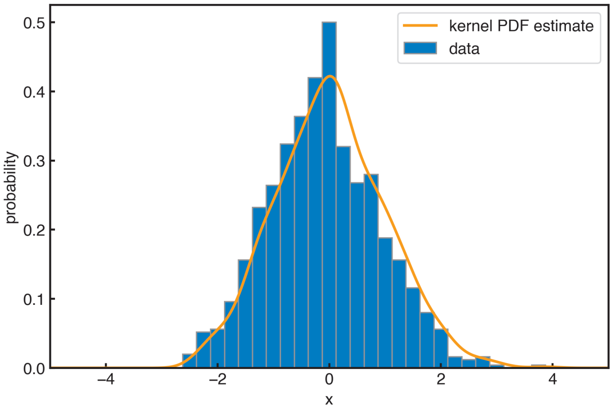

Figure 2.4 Histogram of a data set (blue) along with the kernel estimated probability density function (orange).

2.3.1 The Normal Distribution

The Normal (Gaussian) distribution is one of the most important in nature. Most observables are distributed normally about their means, or can be transformed in such a way that they become normally distributed. Because of this tendency for things to be normally distributed, the most common statistical tests assume normality. Thus, it is very important to verify that your random variable of interest is normally distributed before using common Gaussian statistics.

The probability density function for a normally distributed random variable x is

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} \tag{2.31}$$

The associated cumulative distribution function is

$$F(x) = \int_{-\infty}^{x} \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(t-\mu)^2}{2\sigma^2}} dt \tag{2.32}$$

It is often useful to write the Normal distribution functions in terms of *standardized* random variables, that is, a random variable with mean of 0 and standard deviation of 1. Letting z denote such a standardized random variable,

$$z = \frac{x-\mu}{\sigma}. \tag{2.33}$$

The probability density function and cumulative density function for a Normally distributed, standardized random variable z then simplifies to

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{\frac{-z^2}{2}} \tag{2.34}$$

$$F(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{\frac{-t^2}{2}} dt \tag{2.35}$$

The probability that a standardized, normally distributed random variable z falls within ±1, ±2 and ±3 standard deviations of its mean is given by

$$\boldsymbol{Pr}(-1 \leqslant z \leqslant 1) = \int_{-1}^{1} f(z)dz = 68.27\% \tag{2.36}$$

$$\boldsymbol{Pr}(-2 \leqslant z \leqslant 2) = \int_{-2}^{2} f(z)dz = 95.45\% \tag{2.37}$$

$$\boldsymbol{Pr}(-3 \leqslant z \leqslant 3) = \int_{-3}^{3} f(z)dz = 99.73\% \tag{2.38}$$

These probabilities can also be visualized as the area under the Gaussian $f(x)$ curve, as shown in **Fig. 2.5**. There is only a 4.55% probability that a normally distributed variable will fall more than 2 standard deviations away from its mean. This is a *two-tailed* probability. The probability that a normal variable will exceed its mean by more than 2 standard deviations is only half of that, 2.275%, since the normal distribution is symmetric. This is a *one-tailed* probability.

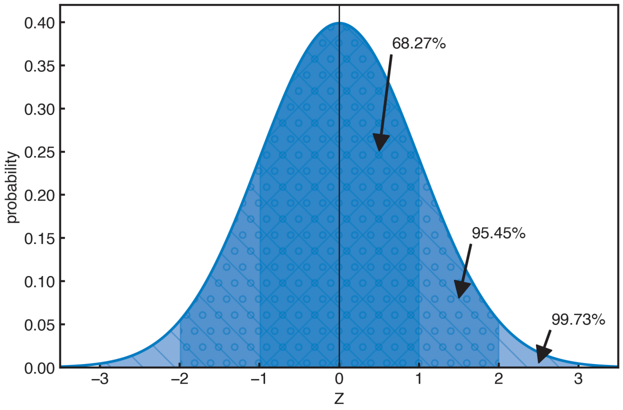

Figure 2.5 Probability density function of z and the probability that z falls within $\pm 1\sigma$, $\pm 2\sigma$ and $\pm 3\sigma$ (area under the curve).

In Practice.

- Standardizing your data using (2.33) comes in handy for comparing particular values with others who may have normally distributed data with different means and standard deviations, or those unfamiliar with the units of your data. For example, if I measured the ozone at a remote site and told you the measurements were normally distributed and today's level was 100 parts per billion (ppb), you may not know what to think. But, if I told you the standardized level was $z = 4\sigma$, you would know that ozone was extremely high today.
- Your data does not need to be Normally distributed to standardize it following (2.33), it is merely a unit conversion, like going from Celsius to Fahrenheit. When this is done, the resulting values can still be interpreted as the number of standard deviations about the sample mean. If the data is not normally distributed, however, the probabilities given in (2.36)-(2.38) and **Fig. 2.5** will not be applicable.

2.3.2 Other Common Distributions

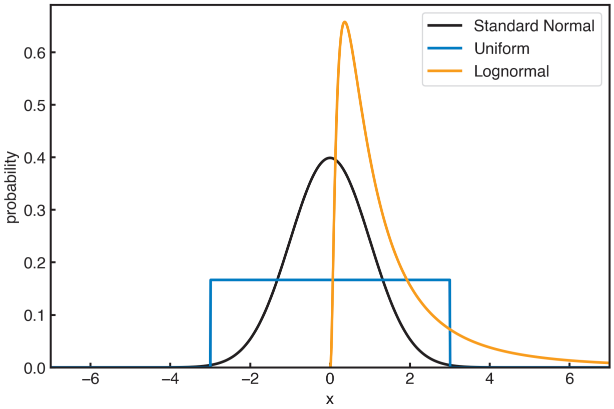

Figure 2.6 The probability density functions of three well-known distributions.

Uniform Distribution

The *uniform distribution* describes a random variable that is equally likely to take any value in the closed interval $[a, b]$. Its probability density function is plotted in **Fig. 2.6** and given by

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{for } a \leqslant x \leqslant b, \\ 0 & \text{for } x < a \text{ or } x > b \end{cases} \tag{2.39}$$

The cumulative distribution function is

$$F(x) = \begin{cases} 0 & \text{for } x < a \\ \frac{x-a}{b-a} & \text{for } a \leqslant x \leqslant b, \\ 1 & \text{for } x > b \end{cases} \tag{2.40}$$

Lognormal Distribution

A positive random variable x has a *lognormal distribution* if the natural logarithm $(\log)$ of x is normally distributed. Put another way, x is lognormally distributed if $Y = \log x$ is normal. To determine the probability density function, it is straight-forward to plug $\log x$ into (2.31), perform a change of variables, and show that the lognormal probability density function is

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} e^{\frac{-(\log x - \mu)^2}{2\sigma^2}}, \; x > 0 \tag{2.41}$$

The cumulative density function is more complicated and requires the complementary error function to be written in full and so we will not do so here. An example lognormal probability density function is shown in **Fig. 2.6**.

Gamma Distribution

The Gamma Distribution is a two-parameter family of distributions. It is included here since it can fit positive-definite highly skewed distributions such as that of precipitation or wind speed. The two parameters are a shape factor, $a > 0$, and a scale factor, $b > 0$. The pdf for the Gamma Distribution is given by

$$f(x) = \frac{1}{\Gamma(a)b^a} x^{(a-1)} e^{-x/b} \tag{2.42}$$

The probability density functions for the gamma distribution with six sets of parameters are shown in **Fig. 2.7**. A small shape and large scale factor give a highly skewed distribution peaking near zero, which can be a good fit to variables like precipitation. A large shape and small scale parameter gives a distribution that is peaked at a non-zero value and less positively skewed.

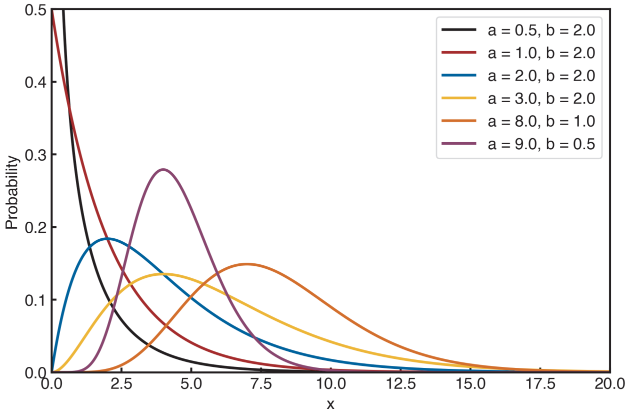

Figure 2.7 The probability density functions for gamma distributions with the shape and scale factors indicated in the legend.

2.4 Central Limit Theorem

Theorem 2.2 (Central Limit Theorem). *The arithmetic mean of a sufficiently large number of iterates of independent random variables, each with a well-defined mean and variance, will be approximately normally distributed, with standard deviation* $\sigma/\sqrt{N}$*, where* N *is the size of each sample.*

In simpler terms, the Central Limit Theorem says that no matter the underlying distribution of your data, if you take a large enough sample of your data, and compute its average, then take another sample and take its average, then another, etc., the distribution of these sample means will be normal with mean equal to the mean of the random variable and standard deviation of $\sigma/\sqrt{N}$, where σ is the standard deviation of the underlying distribution of the random variable. This concept is absolutely fundamental to much of the statistics that we do in the physical sciences, and significantly simplifies the statistics we must master. Let's look at two examples.

Worked Example 2.5.
Suppose we know that our data is normally distributed with $\mu = 0$ and $\sigma = 1$ (standard normal distribution). What is the distribution of sample means for sample sizes of $N = 25$? $N = 100$? $N = 200$?

....

The Central Limit Theorem says that for a "sufficiently large number", that is, for sufficiently large N, the distribution will be normal. But what is "sufficiently large"? It turns out that if the underlying data is normal, the Central Limit Theorem applies for any $N \geqslant 1$. Thus, the sample mean will have a normal distribution with the same mean as the underlying distribution,$\mu = 0$, and standard deviation:

$$\sigma_{N=25} = \sigma/\sqrt{N} = 1/\sqrt{25} = 0.2 \tag{2.43}$$
$$\sigma_{N=100} = \sigma/\sqrt{N} = 1/\sqrt{100} = 0.1 \tag{2.44}$$
$$\sigma_{N=200} = \sigma/\sqrt{N} = 1/\sqrt{200} = 0.07 \tag{2.45}$$

To visualize this result, **Fig. 2.8** displays the distribution of 10000 sample means of values drawn from a standard normal distribution. The dashed gray curves denote the theoretical distribution given in (2.43) - (2.45).

Worked Example 2.6.
In the previous example, the underlying distribution was normal. However, the Central Limit Theorem applies to *all underlying distributions* as long as N is large enough.

Fig. 2.9 shows the distributions of 10000 sample means of length $N = 25, 100, 200$ drawn from the three distributions plotted in **Fig. 2.6**. As in **Fig. 2.8**, the dashed gray curves denote the theoretical normal distribution predicted by the Central Limit Theorem. As N increases, the theoretical estimate and the actual distribution agree more and more. Note how the lognormal distribution of sample means still does not agree completely with the theoretical estimate, and this is also the distribution that is most skewed (looks the least like a Gaussian).

2.5 Testing for Significance

Many geophysical variables are approximately normally distributed, furthermore, as we discussed in Section 2.4, if you take a large enough sample, the sample mean of *any* variable is normally distributed. Thus, we can often use the theoretical normal probability distribution to calculate the probability of measuring a certain value. We have so far covered how to determine the probability of drawing a value x_i within a range of values, but what about comparing a sample's mean to some other value? For example, instead of asking "what is the probability that this summer's average temperature will be greater than 80°F", we might instead want to ask "was this summer's average temperature significantly warmer than that of the summer of 1950?" As in this example, many research questions revolve around determining whether two means are different from

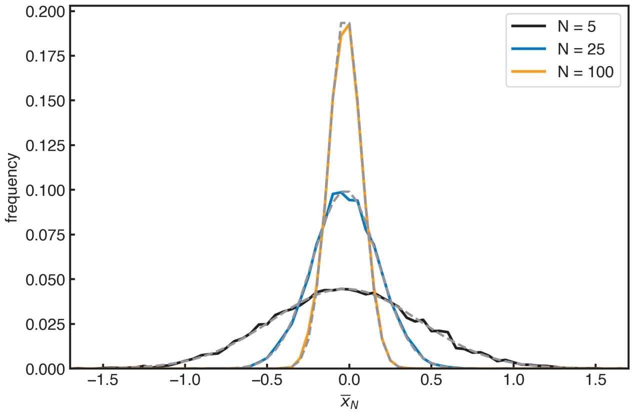

Figure 2.8 Distribution of 10000 sample means drawn from a standard normal distribution for sample sizes of $N = 25, 100, 200$. Dashed gray lines denote the distributions predicted by theory.

one another. To do this we need to know our data's true population mean and population standard deviation *a priori*. Unfortunately, the best that we are likely to have are the sample mean $\overline{x}$ and the sample standard deviation s based on a sample of finite length N.

If we know our data are normally distributed, and N is large enough, then we can use $\overline{x}$ and s to compute the z-statistic. If N is not sufficiently large, we need to use the Student-t distribution (see Section 2.5.1), which is appropriate for small sample sizes.

The standard variable used to compare a sample mean to the true mean is:

$$z = \frac{\overline{x} - \mu}{\sigma_{\overline{x}}} = \frac{\overline{x} - \mu}{\frac{\sigma}{\sqrt{N}}} \tag{2.46}$$

where we have used the Central Limit Theorem to replace $\sigma_{\overline{x}}$ with $\sigma/\sqrt{N}$. The z-statistic is thus the number of standard errors that the sample mean deviates from the true mean. If the variable is normally distributed about its mean, then z can be converted into a probability statement.

(2.46) needs to be altered only slightly to provide a significance test for differences between two sample means:

$$z = \frac{(\overline{x_1} - \overline{x_2}) - \Delta_{1,2}}{\sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}}} \tag{2.47}$$

Here, the sample sizes are allowed to be different, and $\Delta_{1,2}$ is the hypothesized difference between the two means, which is often zero in practice.

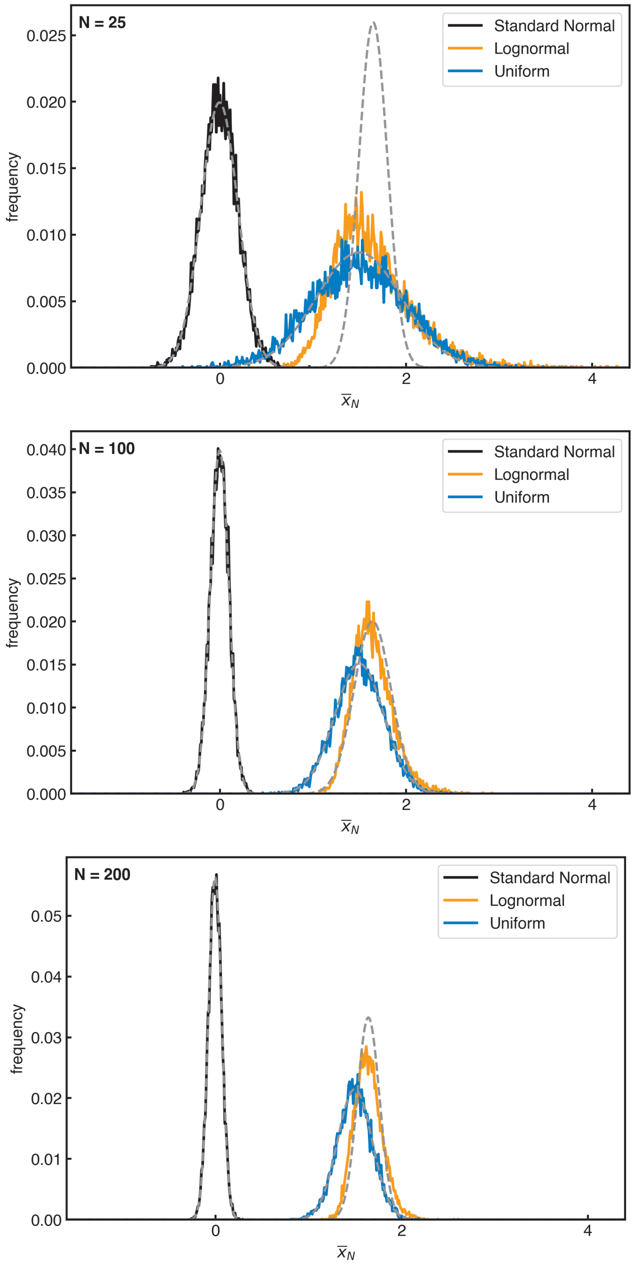

Figure 2.9 Distribution of 10000 sample means drawn from a three distributions for sample sizes of $N = 25, 100, 200$.

2.5.1 Small sampling theory: the t-statistic

When the sample size is smaller than about 30 we cannot use the z-statistic to compare sample means, even if the underlying distribution is normally distributed. Instead, we must use the Student's t distribution to compare sample means, or the chi-squared distribution when comparing sample variances. The key difference between the z-statistic and the t-statistic is that the z-statistic requires knowledge of the population standard deviation σ while the t-statistic uses the sample standard deviation s. When the sample size is smaller than 30, s is biased low as an estimate of σ and thus, we use the t-statistic to account for this.

The Student's t-distribution is derived in exact analogy with the z-statistic:

$$t = \frac{\overline{x} - \mu}{\frac{s}{\sqrt{N-1}}} = \frac{\overline{x} - \mu}{\frac{\widehat{s}}{\sqrt{N}}} \tag{2.48}$$

$$\widehat{s} = s\sqrt{\frac{N}{N-1}} \tag{2.49}$$

If we draw a sample of size N of independent values from a normally distributed population with mean μ, t (as defined by (2.48)) is distributed with the following probability density:

$$f(t) = \frac{f_0(\nu)}{\left(1 + \frac{t^2}{\nu}\right)^{\left(\frac{\nu+1}{2}\right)}}, \tag{2.50}$$

where $f_0(\nu)$ is chosen as a normalization factor to make $\int_{-\infty}^{\infty} f(t)dt = 1$ and $\nu = N - 1$ is the *number of degrees of freedom*. The degrees of freedom is defined as the number of independent samples minus the number of parameters that must be estimated.

In Practice.

- In all cases thus far, it has been assumed that the N values drawn are all *independent* samples. Often, however, N samples of a geophysical variable are not independent, that is, they exhibit either spatial or temporal correlations. For example, geopotential height is highly auto-correlated so that each day's value is not independent from the previous day's. We will discuss how to deal with non-independence in Section 7.4 autocorrelation and degrees of freedom.

Worked Example 2.7.
The Southern Annular Mode (SAM) is the dominant mode of atmospheric variability in the Southern Hemisphere, and can be quantified by a monthly index which is approximately normally distributed with $\mu = 0$ and $\sigma = 1$.

(a) What is the probability that a particular month's SAM index is $\geqslant 0.5$?

(b) What is the probability that the average monthly SAM index over a 4 year period was $\geqslant 0.5$?

....

(a) We are given that, $\mu = 0$ and $\sigma = 1$, and so we can calculate a z-score:

$$z = \frac{\overline{x} - \mu}{\sigma} = \frac{0.5 - 0}{1} = 0.5. \tag{2.51}$$

We want to know $\boldsymbol{Pr}(z > 0.5)$ (i.e. the area under the normal probability density curve that is to the right of 0.5) which can be computed using any software package, and the answer is 31%.

(b) Now, we want to test for the sample mean, with $N = 36$ months. In this case,

$$\overline{x} = 0.5, \sigma_{\overline{x}} = \frac{\sigma}{\sqrt{N}} = \frac{1}{\sqrt{36}} \tag{2.52}$$

$$z = \frac{\overline{x} - \mu_{\overline{X}}}{\sigma_{\overline{x}}} = \frac{0.50 - 0}{.1667} = 3.0 \tag{2.53}$$

The $\boldsymbol{Pr}(z \geqslant 3.0) = 0.1\%$. Such a low probability implies either a very rare event, or, that the dynamics of the SAM over those 4 years was different compared to the climatological SAM variability. Note that in this example we have assumed that each monthly sample is independent, so that the degrees of freedom of the data set equals the number of samples.

Unlike the z-distribution, the t-distribution depends on the size of the sample. The tails of the distribution are longer for smaller degrees of freedom (**Fig. 2.10**). For a large number of degrees of freedom the t-distribution approaches the z or normal distribution. Note that, although we sometimes speak of the t-distribution and contrast it with the normal distribution, the t-distribution is merely the probability density you expect to get when you take a small sample *from a normally distributed population.*

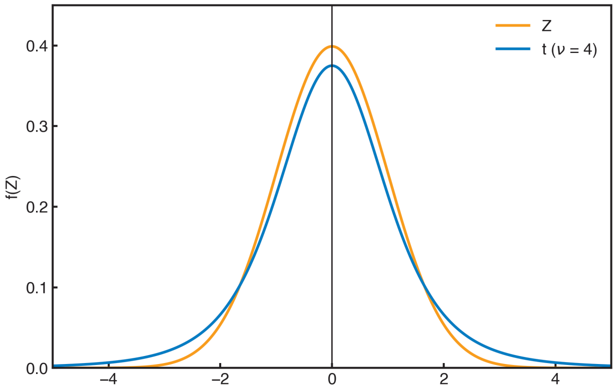

Figure 2.10 Probability density function of $z-$ and t-distribution with $\nu = 4$ degrees of freedom.

In Practice.

- When using the t-statistic, you are making the strong assumption that the underlying distribution is normal. The Central Limit Theorem tells us that for a "large enough" sample size, the distribution of sample means is normal, no matter the distribution. For small sample sizes, the Central Limit Theorem does not apply. Thus, if the underlying population is not normally distributed, and you have a small sample size, you must use other methods.
- Smaller values of N lead to longer tails for the t-statistic, meaning you are more likely to get a sample mean far from the true value when N is smaller.
- Since the t-distribution approaches the normal distribution for large N, there is no theoretical reason to use the z-statistic in preference to the t-statistic, although it maybe be more convenient to do so.

The difference of means for the t-statistic is very similar to that for the z-statistic, but with slight modifications. Assume two samples of length N_1 and N_2 are drawn from an normally distributed population with true standard deviations $\sigma_1 = \sigma_2$, then,

$$t = \frac{(\overline{x_1} - \overline{x_2}) - \Delta_{1,2}}{\widehat{\sigma}\sqrt{\frac{1}{N_1} + \frac{1}{N_2}}} \tag{2.54}$$

$$\widehat{\sigma} = \sqrt{\frac{N_1 s_1^2 + N_2 s_2^2}{N_1 + N_2 - 2}} \tag{2.55}$$

where $\nu = N_1 + N_2 - 2$ and $\Delta_{1,2}$ is the hypothesized difference. The pooled variance $\widehat{\sigma^2}$ is a weighted average of the sample variances.

2.5.2 Confidence intervals

Recall from our discussion of cumulative probability density function F, that

$$\boldsymbol{Pr}(a \leqslant x \leqslant b) = \int_a^b f(x)dx \tag{2.56}$$

$$\boldsymbol{Pr}(a \leqslant x \leqslant b) = F(b) - F(a). \tag{2.57}$$

For a standard normal random variable z, we determined that

$$\boldsymbol{Pr}(-1 \leqslant z \leqslant 1) = 68.27\% \tag{2.58}$$

$$\boldsymbol{Pr}(-2 \leqslant z \leqslant 2) = 95.45\% \tag{2.59}$$

These are confidence intervals for z. The first is the 68.27% confidence interval, and the second is the 95.45% interval.

One can instead first determine a confidence interval of interest, say 95%, and compute the lower-bound a and upper-bound b such that $\boldsymbol{Pr}(a \leqslant z \leqslant b) = 95\%$. Often, the confidence interval is discussed in terms of the parameter α, which is defined as 1 minus the confidence interval. For a 95% confidence interval, $\alpha = 0.05$.

To find the 95% confidence interval for z, or $\alpha = 0.05$, it helps to think in terms of the area under the standard normal probability distribution function (**Fig. 2.11**). That is, we want to find the critical z, denoted z_c, such that

$$\boldsymbol{Pr}\big(z_{c,\alpha/2} \leqslant z \leqslant z_{c,1-\alpha/2}\big) = 0.95. \tag{2.60}$$

Since the normal distribution is symmetric about zero, we can instead write

$$\boldsymbol{Pr}(z) \geqslant z_{c,1-\alpha/2} = 1 - \frac{\alpha}{2} = 0.975. \tag{2.61}$$

We look for 0.975 because we want the total area to add to 5% ($\alpha = 0.05$), and so 2.5% comes from the lower tail and 2.5% comes from the upper tail. Any statistical software can be used to find that for a 95% confidence interval of a normally distributed variable, $z_{c,0.975} = 1.96$.

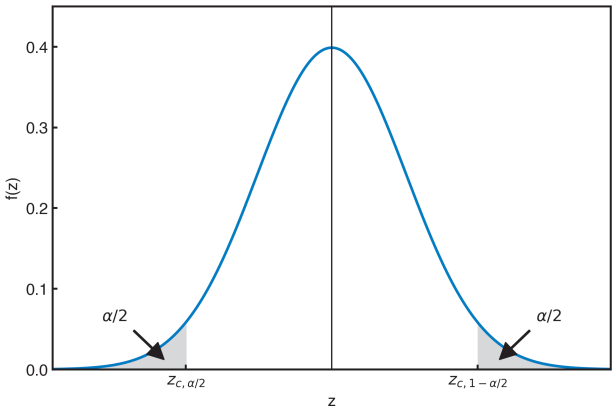

Figure 2.11 Illustration of the relation of the z-statistic probability density function to probability measure α.

The above examples with standardized data, are a relatively straight-forward application of the t-statistic and z-statistic because we are dealing with standardized data. However, what if your data are not standardized? You have two options: you can standardize your data and then do all of your analysis using standard normal variables (as above), or, you can use a modified equation for the confidence interval that takes into consideration the the data's non-zero mean and non-unity standard deviation as we will now demonstrate.

Plugging the definition of the z-statistic (2.33) into (2.60) leads to the 95% confidence interval for any sampled guassian variable x:

$$-z_{c,0.975} \leqslant \frac{x-\mu}{\sigma} \leqslant z_{c,0.975}. \tag{2.62}$$

Following similar steps for the sample mean,

$$-z_{c,0.975} \leqslant \frac{\overline{x}-\mu}{\frac{\sigma}{\sqrt{N}}} \leqslant z_{c,0.975}. \tag{2.63}$$

From this we can deduce that the true mean μ falls within the following bounds 95% of the time:

$$\overline{x} - z_{c,0.975}\frac{\sigma}{\sqrt{N}} \leqslant \mu \leqslant \overline{x} + z_{c,0.975}\frac{\sigma}{\sqrt{N}}. \tag{2.64}$$

In general, confidence limits for population means of symmetric distributions can be represented by

$$\mu = \overline{x} \pm z_{c,1-\alpha/2}\frac{\sigma}{\sqrt{N}} \tag{2.65}$$

Confidence intervals for the sample mean t-statistic are defined similarly,

$$\mu = \overline{x} \pm t_{c,0.975}\frac{s}{\sqrt{N-1}}, \tag{2.66}$$

where t_c is the critical value for t and depends on the significance level desired and the sample size.

Worked Example 2.8.

You have 5 years of monthly-mean temperature data from the MSU4 satellite. The mean temperature around the 60°N latitude circle during January is −60° C and the standard deviation is 8° C. What are the 95% confidence limits on the true population mean? You can assume that monthly-mean temperatures are normally distributed.

....

t-statistic

Since $N = 5$, we must use the t-statistic. The critical value is $t_{c,0.975} = \pm 2.78$ for $\nu = 5 - 1 = 4$. Thus, the population mean μ is expected to lie within

$$-60 \pm 2.78\frac{8}{\sqrt{4}} \Rightarrow -67.0 \leqslant \mu \leqslant -53.0 \tag{2.67}$$

***z*-statistic**

If we had erroneously used the z-statistic, the critical value is $z_{c,0.975} = \pm 1.96$ and the population mean μ would be expected to lie within

$$-60 \pm 1.96\frac{8}{\sqrt{5}} \Rightarrow -71.1 \leqslant \mu \leqslant -48.9 \tag{2.68}$$

Using the t-statistic gives a wider confidence interval than the z-statistic, reflecting the additional uncertainty associated with small N. If we had erroneously used the z-statistic instead of the t-statistic we would have underestimated the 95% confidence bounds by 35%.

2.5.3 Chi-Squared Distribution: Tests of Variance

Sometimes we want to test if the sample variances are truly different. For this we cannot use t-statistic or z-statistic as these are for sample means, but we can use the Chi-Squared distribution. First, define a random variable χ^2:

$$\chi^2 = (N-1)\frac{s^2}{\sigma^2} \tag{2.69}$$

This quantity can be used to test if a sample variance s^2 is different from a population variance σ^2. Note we are using a ratio, rather than a difference.

If the underlying distribution from which we draw N values to compute χ^2 is normally distributed with standard deviation σ, then the χ^2 values themselves will be distributed as follows:

$$f(\chi^2) = f_0(\nu)(\chi^2)^{(\frac{1}{2}\nu-1)} \exp^{-\frac{1}{2}\chi^2}, \quad \nu = N-1 \tag{2.70}$$

where f_0 is a normalization factor. This is the *Chi-Squared distribution* and can be used to estimate the significance of the ratio $\frac{s^2}{\sigma^2}$.

If you wish to determine confidence bounds on the true variance, you can move things around to obtain the confidence limits given your sample variance:

$$\frac{s^2(N-1)}{\chi^2_{c,0.975}} \leqslant \sigma^2 \leqslant \frac{s^2(N-1)}{\chi^2_{c,0.025}}, \quad \nu = N-1. \tag{2.71}$$

Note that the Chi-squared distribution is not symmetric like the normal distribution, and so the lower and upper critical values $\chi^2_{c,0.025}$ and $\chi^2_{c,0.975}$ must both be computed and, like the t-distribution, are functions of the sample size N.

In Practice.

- For $\nu \gtrsim 30$, the Chi-Squared distribution approaches the Normal distribution.

2.6 The Binomial Distribution

2.6.1 Binomial Distribution

Suppose you have a set of N trials in which the outcome is either "success" or "failure". The probability of success in one trial is $p = \boldsymbol{Pr}(\text{success in one trial})$. If X is the total number of successes in N trials, then

$$\boldsymbol{Pr}(X = k) = \binom{N}{k} p^k (1-p)^{N-k} = \frac{N!}{(N-k)!\,k!}\, p^k (1-p)^{N-k}, \quad k = 1, 2, 3 ... N. \tag{2.72}$$

At first, the right-hand-side might look complicated, but note that it is just the probability of k successes times the probability of the rest being failures with an additional factor in front to account for the order of occurrence not mattering.

The binomial distribution is helpful in assessing "field significance", the significance of multiple tests succeeding when an array of variables are tested against the same hypothesis. An example would be correlating the sunspot index with a map of pressure at many points over the earth. How many individual "significant" events do you expect to get by chance in such cases?

As an example, **Fig. 2.12** shows for N tries of a test at the $p = 0.05$ significance level what the binomial distribution (2.72) says about how many you should get by chance alone.

Note that the probability of getting 5 successes or more in 30 tries is less than 0.05 and getting 10 successes or more in 100 tries is less than 0.05. That is 16.7% are successes for 30 tries and only 10% are successes for 100 tries at same probability level. For smaller samples, the fraction of total tries that can succeed by chance is greater. Even for 100 tries, 10% can succeed by chance, where the probability of each individual occurrence is p=0.05. The most likely outcome is shown by the peak of the blue line and is what you expect, about 5% of the chances will succeed. But the chances of getting significantly more than that are quite good, and 10-15% of the field points could succeed by chance at the 5% level (see also Wilks, 2011; Livezey and Chen, 1983; Wilks, 2016).

Worked Example 2.9.

Suppose 14 out of 20 different climate models project that Australia will become drier with increasing greenhouse gas concentrations. What is the probability of getting this result if one assumes that drying and wetting are actually both equally likely under this scenario? That is $\boldsymbol{Pr}(\text{drying}) = \boldsymbol{Pr}(\text{wetting}) = 0.5$?

....

$$\boldsymbol{Pr}(X = 14) = \binom{20}{14} 0.5^{14} (1-0.5)^{20-14} = 0.037 \tag{2.73}$$

What is the probability that 14 *or more* models agree that Australia will become drier if we assume that drying and wetting are both equally likely?

$$\boldsymbol{Pr}(X \geqslant 14) = \sum_{k=14}^{20} \binom{20}{k} 0.5^{k} (1-0.5)^{20-k} = 0.058 \tag{2.74}$$

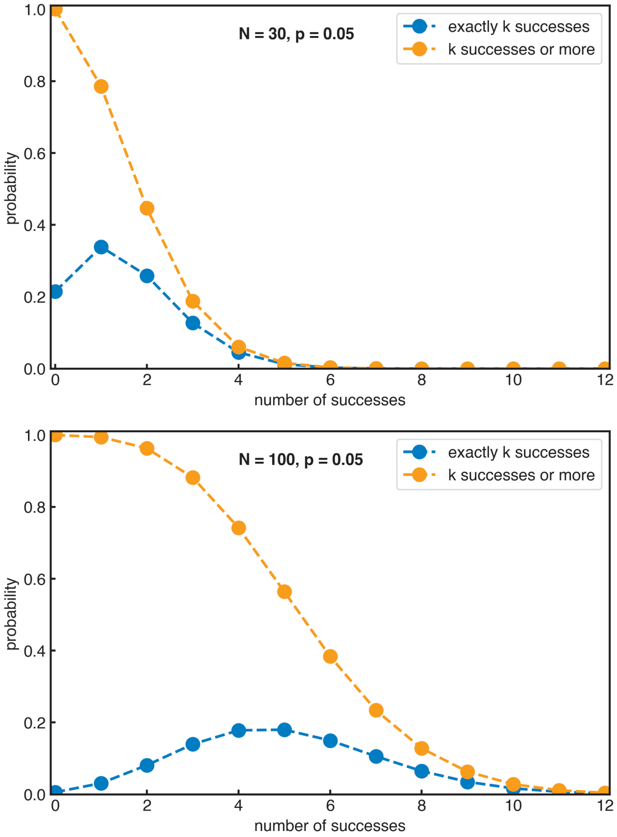

Figure 2.12 Probability of a given number of successes in N trials where the probability of a success is $p = 0.05$.

2.6.2 Normal Approximation to the Binomial

If you did the calculations above by hand you would find it tedious. This gets worse when the sample gets even larger. To assist in this, we can make use of theorem that allows us to use a Normal approximation when performing Binomial calculations.

From the central limit theorem, it can be shown that the distribution of sample means approaches the Normal Distribution, even if the population from which the means are derived is not normally distributed (see Section 2.4). This is also true for the Binomial distribution, for which values have a probability of being either zero or one, but nothing else. The distribution of sample means from a binomial population is nonetheless normally distributed about its mean value of 0.5.

Theorem 2.3 (DeMoivre-Laplace Theorem). *Let X denote a binomial variable defined on N independent trials, each having success probability p. Then, for any numbers a and b,*

$$\lim_{N\to\infty} \boldsymbol{Pr}\left(a < \frac{X - Np}{\sqrt{Np(1-p)}} < b\right) = \frac{1}{\sqrt{2\pi Np(1-p)}} \int_a^b e^{-x^2/2} dx \tag{2.75}$$

This theorem tells us that the statistic $z = \frac{X-Np}{\sqrt{Np(1-p)}}$ follows a normal distribution with $\mu = Np$ and $\sigma = \sqrt{Np(1-p)}$. An approximate two-tailed 95% confidence interval for the number of successes X is then given by

$$Np - 1.96 \cdot \sqrt{Np(1-p)} \leqslant X \leqslant Np_1.96 \cdot \sqrt{Np(1-p)} \tag{2.76}$$

We can use this to simplify the calculation of binomial problems, as illustrated in the examples below.

In Practice.

- When deciding whether a Normal approximation is appropriate to use for your Binomial random variable, some good rules-of-thumb are:
  - large N
  - $Np \geqslant 10$
  - $N(1-p) \geqslant 10$

Worked Example 2.10.
An earthquake forecaster has to forecast 200 earthquakes. How many times in 200 tries must she be successful so we can say with 95% confidence that she has non-zero skill?
....

The null hypothesis is that she has no skill and the significance level is $\alpha = 0.05$. We then want

$$\boldsymbol{Pr}(s > s^*|H_0) = 0.025 = \sum_{s=s^*}^{200} \binom{200}{s} (0.5)^s (1-0.5)^{200-s} \tag{2.77}$$

Solving this equation for $s > s^*$, the number of occurrences necessary to leave only a 0.025 probability to the right, is extremely tedious to do by hand. Instead, we can use the Normal approximation to the Binomial to convert this to the following problem:

$$\boldsymbol{Pr}(s > s^*|H_0) = 0.025 = \boldsymbol{Pr}\left(\frac{s-Np}{\sqrt{Np(1-p)}} > \frac{s^*-Np}{\sqrt{Np(1-p)}}\right) \tag{2.78}$$

$$= \boldsymbol{Pr}\left(Z > \frac{s^*-Np}{\sqrt{Np(1-p}}\right) \tag{2.79}$$

where $\boldsymbol{Pr}(Z > 1.96) = 0.025$ from the standard normal distribution. So, we want

$$\frac{s^*-Np}{\sqrt{Np(1-p}} > 1.96, \quad \text{or} \quad s^* = 114 \tag{2.80}$$

So, to pass a no-skill test on a sample of this size, the forecaster must be right 57% of the time or more.

The 95% confidence interval for the number of successes expected if the forecaster has no skill (i.e. under the null hypothesis) is given by:

$$Np \pm 1.96 \cdot \sqrt{Np(1-0.5)} = \tag{2.81}$$

$$100 \pm 1.96 \cdot \sqrt{10 \cdot 0.5} = \tag{2.82}$$

$$100 \pm 13.86 \tag{2.83}$$

Worked Example 2.11.

Normal Approximation to Binomial: Out of 48 independent climate model simulations, how many must agree that global temperatures will increase by 2100 so that we can say with 95% certainty that the models do not agree purely by chance? What is the 95% confidence interval on the number of models with positive temperature trends under the null hypothesis?

....

Here, let a success be that the model says global temperatures will increase. Our null hypothesis is that the models randomly guess whether global temperatures will increase - thus, there is a 50% chance that any one model will predict a temperature increase ($p = 0.5$). We want to know k^* such that:

$$\boldsymbol{Pr}(X \geqslant k^* | H_0) \leqslant 0.05 \tag{2.84}$$

That is, k^* is the number of models that must show a temperature increase for us to believe it is more than chance (that the null hypothesis can be rejected).

$$\sum_{k=k^*}^{48} \binom{48}{k} (0.5)^k (1-0.5)^{48-k} \leqslant 0.05 \tag{2.85}$$

This would take a long time by hand, however, we can instead use the Normal approximation to the Binomial:

$$\begin{align} \boldsymbol{Pr}\left(Z > \frac{k^* - 48 \cdot 0.5}{\sqrt{48 \cdot 0.5 \cdot (1-0.5)}}\right) &= 0.025 \tag{2.86}\\ \frac{k^* - 48 \cdot 0.5}{\sqrt{48 \cdot 0.5 \cdot (1-0.5)}} &= 1.96 \tag{2.87}\\ k^* &\geqslant 31. \tag{2.88} \end{align}$$

So, at least 31 models must show increasing temperatures to reject the null hypothesis that the model agreement in a warming trend is due to random chance. As expected, more than half of the models must show an increase.

The 95% confidence interval under the null hypothesis is:

$$\begin{align} Np \pm 1.96 \cdot \sqrt{Np(1-p)} \tag{2.89}\\ 24 \pm 1.96 \cdot \sqrt{24(1-.5)} = 24 \pm 7 \tag{2.90} \end{align}$$

2.7 The Poisson Distribution

The Poisson distribution applies when you are counting the number of objects in a certain interval. The interval can be in space (volume, area or length) or time. You know the average number of counts per unit interval, and wish to know the chance of actually observing various numbers of objects or events. We denote the associated random variable N, since they are actual counts.

$$N \Rightarrow \text{Poisson}(\lambda) \tag{2.91}$$

There are three necessary and sufficient conditions for a Poisson Distribution.

1. Two or more events cannot occur simultaneously. This means that the events themselves occupy negligible space (e.g. volume, area, length, time).

2. Events occur at an average rate of λ (per unit e.g. volume, area, length, time). This means that λ cannot be a function of space or time.
3. Events occur independently (i.e. they do not know about each other)

The probability mass function of a Poission is defined by the probability that $N = n$ in a given interval of magnitude t according to:

$$\boldsymbol{Pr}(N = n) = \frac{(\lambda t)^n}{n!} e^{-\lambda t} \qquad \lambda > 0, t > 0, \text{ and } n = 0, 1, \ldots \tag{2.92}$$

The first and second moments (mean and variance) are given by:

$$\mu = \lambda t \qquad \sigma^2 = \lambda t \tag{2.93}$$

Estimating $\widehat{\lambda}$ from your data is quite straightforward. Let N be the number of observed events in time t, and assume that N is well-modelled as a Poisson Distribution with unknown rate parameter λ (where λ has units of "events per unit time"). The rather obvious formula for estimating the rate parameter is then simply the number of events divided by the time over which they were observed:

$$\widehat{\lambda} = \frac{N}{t} \tag{2.94}$$

The standard deviation (or standard error) of this estimator is

$$\sigma_{\widehat{\lambda}} = \sqrt{\frac{\lambda}{t}} \tag{2.95}$$

Putting these together, the approximate confidence interval for the true parameter λ (assuming the Central Limit Theorem applies, which it does for $N > 30$ or so) is given by

$$\boldsymbol{Pr}\Big(\widehat{\lambda} - z_{\alpha/2}\sigma_{\widehat{\lambda}} \leqslant \lambda \leqslant \widehat{\lambda} + z_{\alpha/2}\sigma_{\widehat{\lambda}}\Big) \tag{2.96}$$

Worked Example 2.12.
Poisson rate confidence interval: Let's say we count 137 events in 44 minutes. Our estimated rate parameter is

$$\widehat{\lambda} = \frac{N}{t} = \frac{137}{44} \approx 3.11 \text{ events per minute} \tag{2.97}$$

The approximate standard error for the estimated rate parameter is

$$\sigma_{\widehat{\lambda}} = \frac{\sqrt{N}}{t} \frac{\sqrt{137}}{44} \approx 0.27 \text{ events per minute} \tag{2.98}$$

and so the approximate 95% confidence interval for the true, but unknown, rate parameter is

$$\widehat{\lambda} - 1.96\sigma_{\widehat{\lambda}} \leqslant \lambda \leqslant \widehat{\lambda} + 1.96\sigma_{\widehat{\lambda}} \Rightarrow 2.58 \leqslant \lambda \leqslant 3.64 \tag{2.99}$$

It turns out that the Poisson Distribution has a close relationship with the Binomial Distribution. That is, for $n \to \infty$, $p \to 0$, with $np \to \lambda \neq 0$, the Binomial Distribution converges to the Poisson Distribution with parameter λ. In practice, the Binomial Distribution may be approximated by the Poisson when $p < 0.5$ and $n > 20$.

One might be interested in whether the Poisson rates in two samples are different or not. Suppose we have two rates λ_1 and λ_2 drawn from samples of size t_1 and t_2. Our null hypothesis is that $\lambda_1 - \lambda_2 = 0$. The pooled-rate test is based on a standard normal statistic defined as follows.

$$z = \frac{\lambda_1 - \lambda_2}{\sqrt{\widehat{\lambda}\left(\frac{1}{t_1} + \frac{1}{t_2}\right)}} \tag{2.100}$$

where

$$\widehat{\lambda} = \frac{t_1\lambda_1 + t_2\lambda_2}{t_1 + t_2} \tag{2.101}$$

Worked Example 2.13.
Poisson rate difference test: According to the ERA reanalysis data set, during the 28 years from 1952 to 1979 there were 14 major stratospheric warmings, and during the 42 years from 1980 to 2021 there were 25 warmings. The rates are thus 0.05 per year and 0.595 per year. Are these rates different at p=0.05?

$$\widehat{\lambda} = \frac{42 * 0.595 + 28 * 0.5}{42 + 28} = 0.557 \tag{2.102}$$

$$z = \frac{0.595 - 0.5}{\sqrt{0.557\left(\frac{1}{42} + \frac{1}{28}\right)}} = 0.52 \tag{2.103}$$

This is much less than the critical z value of 1.96 for a two-tailed test, so we cannot reject the null hypothesis that the rates of occurrence are the same for the two intervals.

2.8 Non-parametric Statistical Tests

The statistical tests applied above mostly assume that the samples come from populations for which the statistical distributions are known, or assumed, *a priori*. We very often assume that the statistics we are testing are Normally distributed, so we can use the shape of the Normal distribution in our tests. Tests have also been developed that do not require the assumption of a theoretical distribution. These are called *non-parametric* or *distribution-free* statistical tests.

2.8.1 Signs Test

Suppose we have paired data (x_i, y_i) and we want to know if the mean of x_i is different from the mean of y_i. By *paired data*, we mean that each x_i is uniquely associated with a y_i. If we have a suspicion that the data are not normally distributed, and we do not have enough data to invoke the Central Limit Theorem, we cannot use the t-test or the z-test. Instead, if we formulate our question in terms of the median $(\widetilde{\mu})$, rather than the mean, our null hypothesis is that the medians of x_i and y_i are the same, and the alternative is that they are not:

$$H_0 : \widetilde{\mu}_1 = \widetilde{\mu}_2 \quad H_1 : \widetilde{\mu}_1 \neq \widetilde{\mu}_2 \tag{2.104}$$

Let's reformulate this in terms of a probability that y_i is greater than x_i (noting that we could as easily formulate it as less than)

$$H_0 : \boldsymbol{Pr}(y_i > x_i) = 0.5 \quad H_1 : \boldsymbol{Pr}(y_i > x_i) \neq 0.5 \tag{2.105}$$

To test this null hypothesis, we can use the Signs Test. To perform this test, we simply replace each (x, y) pair with a signed integer equal to 1 according to the following rule:

$$y_i > x_i \rightarrow +1 \tag{2.106}$$
$$y_i < x_i \rightarrow -1 \tag{2.107}$$
$$\tag{2.108}$$

The null hypothesis would suggest that there will be a similar number of positive and negative ones (both are equally probable). With this setup we now have a set of Bernoulli trials with success (+1) and failure (−1). We know that the number of successes over N trials will be binomially distributed, and so we can use this distribution to determine whether our actual success rate is outside of what might be expected given our null hypothesis.

Worked Example 2.14.

Cloud Seeding Experiment: Ten pairs of very similar developing cumulus clouds were identified. One from each pair was seeded, and the other was not. Then the precipitation falling from the clouds later was measured with a radar. The data in the following table resulted:

Cloud Pair	x_i: Precip. (untreated)	y_i: Precip. (treated)	$y_i > x_i$?
1	10	12	+1
2	6	8	+1
3	48	10	−1
4	3	7	+1
5	5	6	+1
6	52	4	−1
7	12	14	+1
8	2	8	+1
9	17	29	+1
10	8	9	+1

....

Using the data above, we get 8 +1 and 2 −1. Are these results inconsistent with the null hypothesis that cloud seeding has no effect on precipitation? Can we confidently say that the median values of the two samples are different at 95% confidence? We can plug our values into the binomial distribution to determine the probability of getting 8 successes in 10 tried.

$$\boldsymbol{Pr}(k \geqslant 8) = \sum_{k=8}^{10} \binom{10}{k} 0.5^k (1-0.5)^{10-k} = 0.055 \tag{2.109}$$

If things are random (our null hypothesis is true), the chance of getting two or fewer successes is equally probable as getting 8 or more. Using a two-sided test we find that the probability our result is $p = 0.11$, which fails a 95% confidence test. We expect to toss 8 out of ten heads or tails about 11% of the time.

The Signs Test is one of the simplest non-parametric tests available, but it also has its limitations. For example, it will not tell you the magnitude of the difference of the medians. There are many other distribution-free tests that can be used, for example, the *Wilcoxon signed rank test* and the *Wilcoxon-Mann-Whitney test*.

2.8.2 Rank Sum Test

Another common and classical non-parametric test is the *Rank-Sum Test* (or Wilcoxon-Mann-Whitney Test). Suppose we have two samples S_1 and S_2 of sizes N_1 and N_2 and we wish to test the null hypothesis that they both were sampled from the same distribution (whatever it is). The first step to the Rank-Sum Test is to combine them into a single sample $N = N_1 + N_2$ and rank them from smallest (rank $r = 1$) to largest rank $r = N$). Next, compute the sum of the ranks of each sample S_1 and S_2 and call these R_1 and R_2.

R_1/N_1 and R_2/N_2 should be similar if our null hypothesis is true and the two samples are from the same underlying distribution. Thinking a little harder, one can see that there are $N!/(N_1!N_2!)$ possible combinations of R_1 and R_2. Mann-Whitney showed that the U statistic could be used to determine the probability of a particular combination where

$$U_1 = R_1 - \frac{N_1}{2}(N_1 + 1) \tag{2.110}$$

$$U_2 = R_2 - \frac{N_2}{2}(N_2 + 1) \tag{2.111}$$

where

$$U_1 + U_2 = \frac{N_1 N_2}{2}. \tag{2.112}$$

The U statistic is approximately Normally distributed with mean and standard deviation

$$\mu = \frac{N_1 N_2}{2} \tag{2.113}$$

$$\sigma = \left(\frac{N_1 N_2 (N_1 + N_2 + 1)}{12} \right)^{1/2} \tag{2.114}$$

The statistical significance of U can then be tested with the standard z-score.

2.8.3 Runs Test (Wald-Wolfowitz Test)

The Runs Test is a non-parametric test to check whether a list of values is random or not. For example, imagine a time series of anomalies as shown below, where "+" denotes a positive anomaly and "-" denotes a negative anomaly:

$$\underbrace{++++}_{\text{Run 1}}\underbrace{---}_{\text{Run 2}}\underbrace{+++}_{\text{Run 3}}\ \underbrace{--}_{\text{Run 4}}\ \underbrace{++++++}_{\text{Run 5}}\underbrace{----}_{\text{Run 6}} \tag{2.115}$$

We now separate this series into *runs*; there are a total of $R = 6$ runs, three of which are runs of "+" and three of which are runs of "-". The Runs Test tests the null hypothesis that the data set is random. Under this null hypothesis, the number of runs (R) in a sequence of N elements is a random variable whose conditional distribution given the observation of N_+ positive values and N_- negative values ($N = N_+ + N_-$) has the following properties:

$$\mu = 1 + \frac{2N_+N_-}{N} \tag{2.116}$$

$$\sigma^2 = \frac{2N_+N_-(2N_+N_- - N)}{N^2(N-1)} = \frac{(\mu - 1)(\mu - 2)}{N - 1} \tag{2.117}$$

If N_+ and N_- are each sufficiently large (say, each greater than 30) then the number of runs R is well modeled by a Normal distribution with parameters μ and σ given above. One can then use a typical z-score to determine the probability of obtaining the number of observed runs R under the null hypothesis that the data set is random.

2.8.4 Kolmogorov-Smirnov Test

The Kolmogorov-Smirnov Test (or *KS Test*) tests the equality of two continuous, one-dimensional probability distributions. The most standard version tests whether a particular sample distribution is the same as a specific reference distribution. Because it is a non-parametric test, you do not need to know what the true distribution of your data is, however, the test will not tell you what distribution your data follows either. It will only give you information about its similarity to another reference distribution. Finally, the KS Test is sensitive to both *location* and *shape* and thus cannot tell you why the distributions are different (e.g. is the sample distribution shifted compared to the reference or is the sample distribution wider than the reference?).

The KS Test works by comparing the cumulative density functions (CDFs) of a sample of length N and a reference distribution. Specifically, the difference between the two CDFs is computed, and the *maximum difference*, denoted as D, is used as the test statistic. The null hypothesis is rejected at the significance level α if

$$\sqrt{N}D > K_\alpha \tag{2.118}$$

where K_α is defined as

$$\boldsymbol{Pr}(K \leqslant K_\alpha = 1 - \alpha) \tag{2.119}$$

and the probability density function of K is defined as

$$\boldsymbol{Pr}(K \leqslant x) = 1 - 2\sum_{j=1}^{\infty}(-1)^{j-1}e^{-2j^2x^2} \tag{2.120}$$

If you are specifically interested in whether your sample distribution is normal, other tests may be better suited (e.g. the Shapiro-Wilks or the Anderson-Darling test). In addition, it is important that you do not estimate the parameters of the *reference distribution* from the data. The test is not valid if you do. Thus, if you are comparing to a normal distribution and don't know the true mean and standard deviation of your sample population, you should standardize your data first and compare the standardized sample to the standard normal. Finally, note that the above discussion only applies when you wish to compare a single sample to some reference distribution. What if instead you wish to compare two sample distributions? For that, you can use the *Two Sample KS-Test* which is similar to the standard KS Test and will not be discussed in detail here.

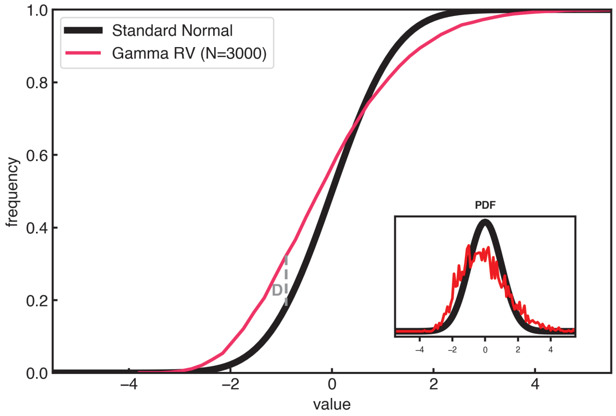

Figure 2.13 Comparison of the standard normal with a random variable drawn from a Gamma distribution. D is the maximum difference between the sample CDF and the reference (normal) CDF and is the test statistic used by the KS-test.

2.9 Hypothesis Testing

2.9.1 Terminology and symbology

- **significance level** [α]: the probability of a false positive (Type I error), often reported as $(1 - \alpha)\%$ confidence level
- **critical value** [t_c, z_c]: the value that must be exceeded to reject the null hypothesis using a significance level of α
- **p-value**: the probability of observing an effect given that the null hypothesis is true (probability of the actual statistic you calculated from your data given the null hypothesis is true)

2.9.2 Setting-up the problem

Hypothesis testing involves stating a hypothesis (*null hypothesis*), and then computing statistics to quantify the extent to which your data set is (or is not) consistent with this hypothesis. The significance level (α) of a hypothesis test defines the probability of a false positive (i.e. Type I error), that is, stating that your data set is not consistent with the null hypothesis when in fact it is. This significance level is a choice that should be made by the scientist.

When performing a statistical hypothesis test there are five basic steps that should be followed in order:

1. State the significance level (α)
2. State the null hypothesis H_0 and the alternative H_1
3. State the statistic of interest
4. State the critical region
5. Evaluate the statistic and state the conclusion

Proper construction of the null hypothesis and its alternative is critical to the meaning of statistical significance testing. Careful logic must be employed to ensure that the null hypothesis is reasonable and that its rejection leads uniquely to its alternative. Usually the null hypothesis is a rigorous statement of the conventional wisdom or a “zero information conclusion”, and its alternative is an interesting conclusion that follows directly and uniquely from the rejection of the null hypothesis. Usually the null hypothesis and its alternative are mutually exclusive.

Worked Example 2.15.
Examples of null hypotheses and their alternatives:

H_0: the means of two samples are equal

H_1: the means of two samples are not equal

H_0: the correlation coefficient is zero

H_1: the correlation coefficient is not zero

In Practice.

- Hypothesis testing tends to yield weak statements. All you can do is state whether or not the data are consistent with the null hypothesis. You cannot state whether the null hypothesis is true or whether the alternative hypothesis is true, or even whether either is false.

2.9.3 Type I and Type II errors in hypothesis testing

Even though you have applied a test and the test gives you a result, you can still be wrong. The following table illustrates the two different types of errors that can be made:

- **Type I:** reject the null hypothesis when it is actually true
- **Type II:** fail to reject the null hypothesis when it is actually false

	H_0 is true	H_0 is false
Fail to Reject H_0	No Error	Type II Error
Reject H_0	Type I Error	No Error

The way typical hypothesis tests are set up, a 95% confidence level means you have a 5% chance of making a *Type I Error*, that is, you reject the null hypothesis (e.g. think you found something interesting) when you should not have. It is much more difficult to asses the Type II Error - the probability you "play it safe and fail to reject H_0 when something interesting was there". For typical hypothesis testing, the probability of a Type II error can be very large.

In Practice.

- One often cares about the differences between probabilities of Type I and Type II errors. For example, if H_0 is that the bridge will hold-up if 10 semi-trucks cross at the same time, and H_1 is that the bridge will not hold-up, you might be happier with a Type I Error, which requires that you redesign the bridge, rather than a Type II Error, where you think the bridge will be fine, and it won't be.
- When performing a hypothesis test, it is good practice to determine α before performing any calculations. But which α should you choose? The choice of α depends on your risk tolerance, that is, the risk you are willing to take to have a Type I error - the smaller the α, the lower the risk. In atmospheric science, α is typically equal to 0.05, 0.01 or sometimes, 0.10, but it is up to the scientist to decide which value of α is best for the hypothesis being tested.

2.9.4 a priori vs. a posteriori

When performing hypothesis tests it is critical to make the distinction between *a priori* and *a posterior* information.

- ***a priori*:** you have reason to expect a particular relationship ahead of time
- ***a posteriori*:** you don't.

One place where such a distinction arises is whether to use a one-tailed or two-tailed hypothesis test. If you have an *a priori* expectation of the tail of interest, you can use a one-tailed test. Otherwise, you should use a two-tailed test. Since your *a priori* expectation might be regarded as subjective by another scientist, it is generally a better practice to use a two-tailed test.

Another common example is when the same hypothesis test is run many times for similar data sets, for example, testing the significance of anomalies at every grid point on the globe. If one does not take into consideration that the test was run hundreds, if not thousands, of times, they will likely be misled thinking there are more significant anomalies than there really are. In effect, you may be giving your hypothesis many chances to succeed. These concepts are perhaps best illustrated with examples (see below).

Worked Example 2.16.

***a priori vs a posteriori*:** You think that climate change has caused the frequency of severe weather to increase between the 1980's and today. You divide the globe into 20 regions and within each region analyze data for each of the 4 seasons. You test for changes in severe weather frequency using $\alpha = 0.05$ (95% confidence level). How many "significant" changes should you expect by chance alone? How might you apply *a posteriori* statistics?

....

You have no *a priori* reason to expect a particular region or season should exhibit changes due to climate change, so you test them all. That is, you perform $N = 4 \times 20 = 80$ different hypothesis tests with the null hypothesis H_0 : the frequency of extreme weather has not changed. By chance alone, you expect on average 5% of these tests to reject the null hypothesis when it is in fact true, or, you expect 4 region/season combinations to produce "significant" changes, purely by chance.

$$\boldsymbol{Pr}(\text{correctly not reject } H_0 \text{ when it is true for one test}) = 0.95$$
$$\boldsymbol{Pr}(\text{correctly not reject } H_0 \text{ when it is true for all 80 test}) = 0.95^{80} \approx 1.7\%$$

Thus, your 95% confidence level is really a 1.7% confidence level! In other words, you have a 98.3% chance of finding at least one significant change, even if climate change has no impact.

Using *a posterior* statistics, we can instead calculate the significance level α for which $\alpha^{80} \approx 0.95$. In this case, $alpha \approx 0.9994$. Thus, if we require that severe weather frequency changes for each region/season combination pass at the 99.94% confidence level, the probability of correctly not rejecting the null hypothesis for all 80 chances will be 95%. We can also use the Binomial Distribution to assess the likelihood of getting some number of "significant" changes above the expected value of 4.

2.9.5 Field significance and False Discovery Rate

Much of geophysical research involves creating maps of a result, and often, scientists will assess the significance of each value on the map individually. As discussed above, one should expect a certain fraction of points to be "significant", even if the null hypothesis is true. Furthermore, many geophysical variables are spatially correlated, implying that significant points will likely appear clustered. An example of this is illustrated in **Fig. 2.14**. To create this figure, daily January 500 hPa geopotential heights at each latitude/longitude grid point is correlated with a time series X. Correlations different from zero at 95% confidence are stippled, and appear to show signals across the globe, with the largest signal in the tropical Pacific. The trick here is that X is a random Gaussian time series, with absolutely no physical meaning, and yet a large cluster of data points were found to be significantly correlated. Thus, in many applications, assessing the significance at each grid point is not enough - rather - one should assess the collective significance, or *field significance* over the entire map (Livezey and Chen, 1983).

Wilks (2016) outlined a straight-forward way to assess field significance by controlling the *false discovery rate* (FDR), or, the expected rate of rejected local null hypotheses where the actual null hypothesis is true. The general idea is that one sorts the list of p-values (across grid points), and then finds which p-value intersects the line defined by

$$y = \frac{i}{N}\alpha_{FDR} \tag{2.121}$$

where i is the position of the p-value in the sorted list, N is the total number of grid points, and α_{FDR} is a parameter that is chosen by the user. The p-value at the intersection is then the global p-value that each grid point must be smaller than to satisfy a particular false detection rate. An illustration of the calculation of this global p-value threshold is shown in the left panel of **Fig. 2.15** for the example plotted in the bottom panel of **Fig. 2.14**. There is no intersection of the actual p-values with the FDR criterion line given in **(2.121)**, and so, none of the stippled points in **Fig. 2.14** should be considered globally significant. For a case where we expect a physical relationship, that is, the correlation of daily January 500 hPa geopotential heights with the stratospheric zonal winds, an intersection occurs and the new global $p-value$ threshold is actually 0.066 rather than 0.05 (right panel of **Fig. 2.15**).

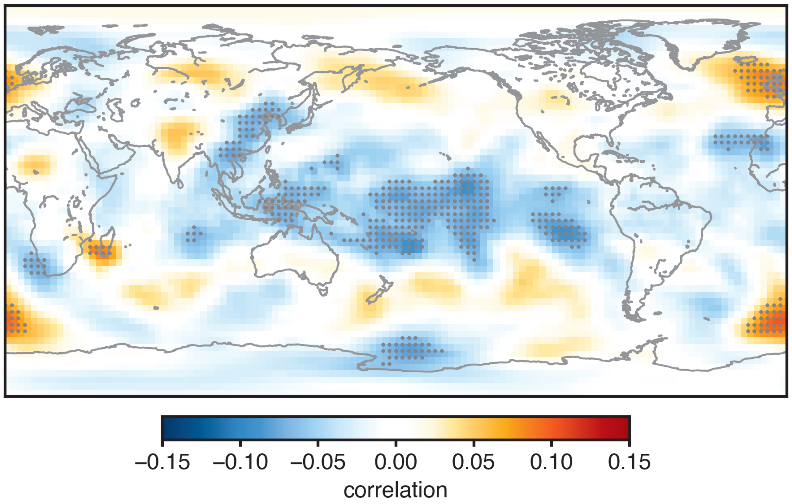

Figure 2.14 Correlation of daily January 500 hPa geopotential heights (1979-2011) with a random Gaussian time series. Statistically significant correlations at 95% confidence ($\alpha = 0.05$).

2.10 Extreme Value Theory

Extreme events are those that appear in the tails of the probability distribution (Coles, 2001). While rare, they can have very important impacts, and so understanding their frequency is very important. Design of physical and financial infrastructure must take into account the most extreme events that are likely to occur over some defined period of time. Therefore the study of extreme values is very important, particularly during this time of global change.

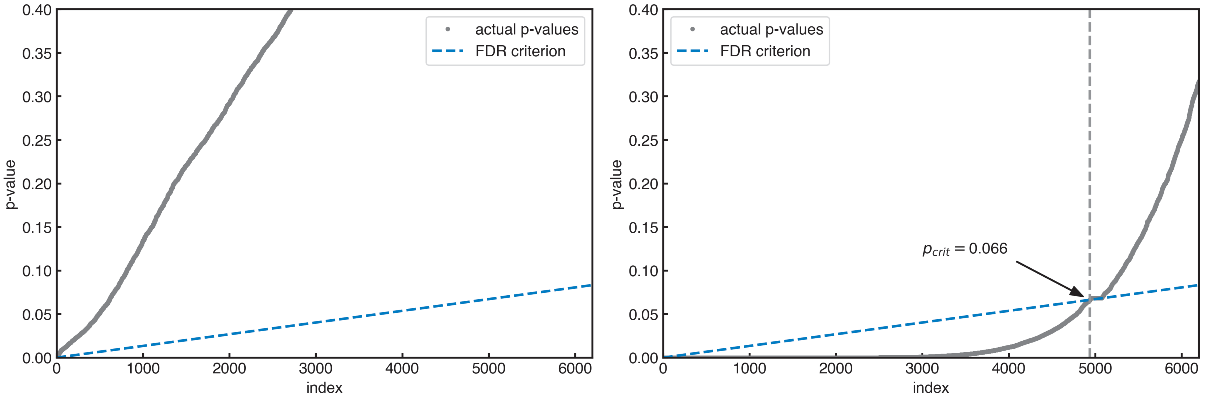

Figure 2.15 Illustration of the false detection rate criterion of Wilks (2016) for (left) correlations with a random time series as shown in **Fig. 2.14**, (right) correlations with daily January 100 hPa polar cap zonal winds averaged around the 65° latitude circle. In both panels $\alpha_{FDR} = 0.1$.

2.10.1 Fisher-Tippett Theorem and Generalized Extreme Value Distribution

Suppose we have a sample of n independent and identically-distributed random variables $[X_1, X_2, X_3, .., X_n]$, each of which has the same cumulative distribution function F. Suppose further that there exists two sequences of numbers $a_n > 0$ and $b_n \in \Re$ such that the following limits converge to a non-degenerate distribution function $G(x)$.

$$\lim_{n\to\infty} \boldsymbol{Pr}\left(\frac{\max\{X_1, \ldots, X_n\} - b_n}{a_n} \leqslant x\right) = G(x) \tag{2.122}$$

which is equivalent to

$$\lim_{n\to\infty} (F(a_n x + b_n))^n = G(x) \tag{2.123}$$

The Generalized Extreme Value (GEV) distribution has three parameters; location $= \mu$, scale $= \sigma$ and shape $= \xi$. Using the definition $s = (x - \mu)/\sigma$ the pdf of the GEV distribution is given by,

$$f(\, x \,|\, \mu, \sigma, \xi) = \frac{1}{\sigma}(1 + \xi\, s)^{-\frac{1}{\xi} - 1} \exp[-(1 + \xi\, s)] \tag{2.124}$$

For particular values of ξ the GEV divides into the Gumbel, Fréchet and Weibull families of distributions, corresponding to the cases $\xi = 0$, $\xi > 0$ and $\xi < 0$, respectively . Each of these distributions has a range of x in which they are supported, which depends on the shape and scale.

In addition, the Generalized Pareto Distribution is often used to describe extreme values. The pdf of the Pareto distribution is given by,

$$f(\, x \,|\, \sigma, \xi) = \frac{1}{\sigma}(1 + \frac{\xi\, x}{\sigma})^{-\frac{1}{\xi} - 1} \quad x \in (0, \infty) \tag{2.125}$$

Dennis: *Need more detail and examples here. Finish project on Pareto fit to SeaTac Max temps*

2.11 Monte Carlo and Resampling

2.11.1 Monte Carlo Techniques

In the age of computers, sometimes it is easier to let the computer do the work by performing many intelligently designed calculations, and then infer a fact or statistical conclusion from the aggregate of these calculations. These techniques take advantage of the computational power at our fingertips and are incredibly powerful when data size is not an issue. The name *Monte Carlo* comes from the famous casino, not from the inventor of the method. It is a term that has no precise definition and covers a wide variety of techniques, which share in common the idea expressed in the first sentence.

One famous example is the calculation of π - the ratio of the circumference of a circle to its diameter. Rather than trying to derive it from basic principles, one can instead write a simple computer code to get a very accurate approximation. Specifically, π can be calculated by inscribing a circle within a square, dropping pebbles randomly on the square, and then counting the ratio of the pebbles in the square to those that fall within the circle. If the pebbles are dropped randomly, then this ratio should be the ratio of the areas of the circle to the square, which is $\pi/4$. If you do this many times you can get an arbitrarily good approximation to π.

2.11.2 Resampling via Bootstrapping

Bootstrap Resampling involves constructing a number of random *resamples* of a dataset of equal size to the true sample of interest. In this way, you do not need to assume anything about the underlying distribution of the data since it is already built into the original dataset. In essence, you ask, by random chance, what is the probability that a particular event (or sample statistic) occurred?

This method is also useful when you are determining statistics other than the mean (e.g. extrema, median, skewness) when we don't have simple statistics for these variables.

The advantage of this method is that you don't have to choose a model PDF and you can evaluate the number of successes in exceeding the criteria using the binomial distribution.

A question arises of whether one should perform the random sampling *with* or *without replacement.* Namely, should you be able to pick the same value twice for the same random sample? Most often, bootstrapping resampling is done *with replacement* as the data set used for sampling is meant to represent an entire population of possibilities. More practically, if your data set is large enough, with and without replacement should give nearly identical answers.

The technique is called "bootstrapping" because it almost appears you get something out without putting something in. This is a phrase often used in computer programming to refer to a small amount of simple software that can load more complex software that loads more complex software (etc., etc.,) almost as if the program is "pulling itself up by its bootstraps."

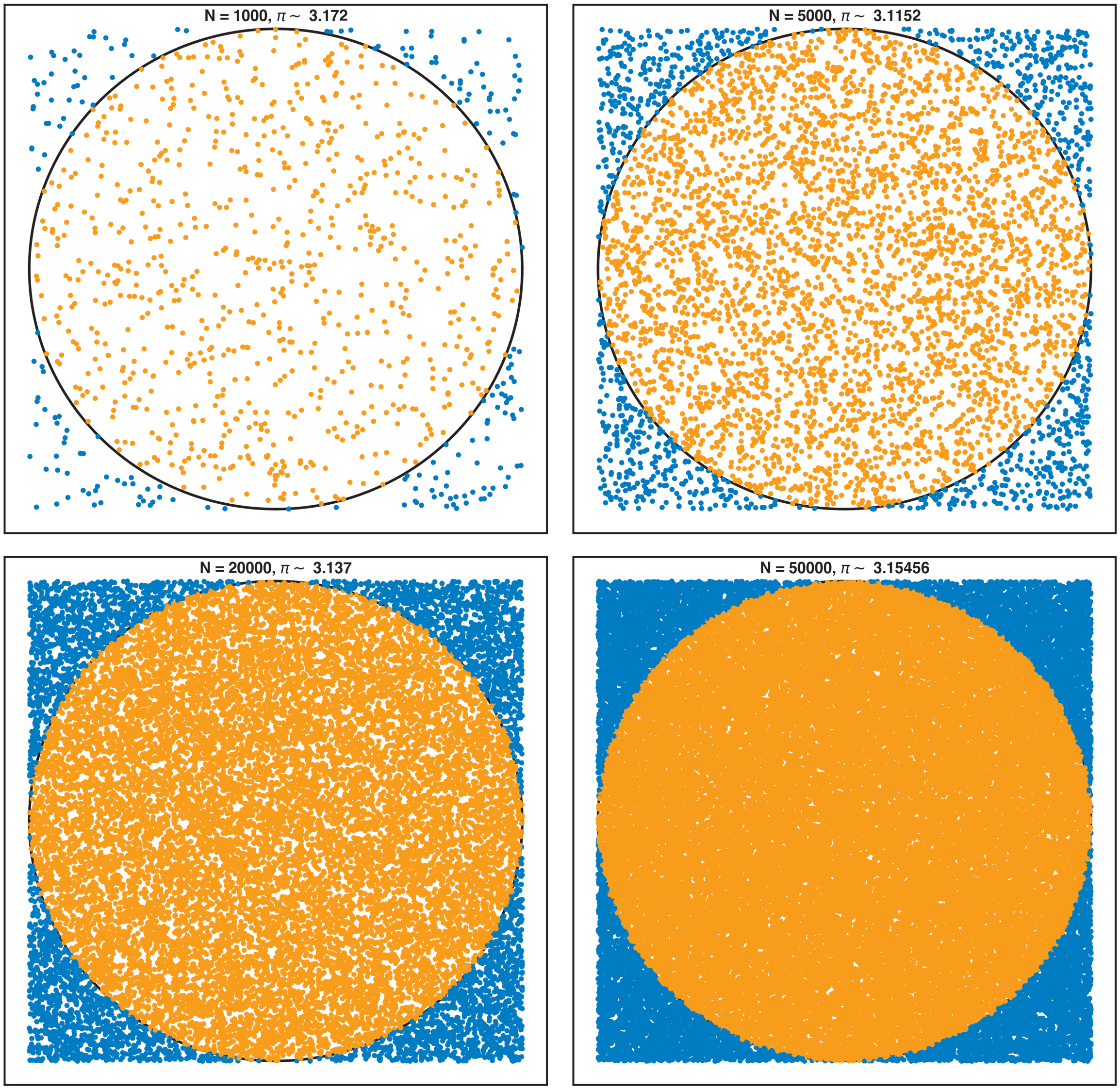

Figure 2.16 Estimate of π using a Monte Carlo approach.

Worked Example 2.17.
A frequently used application of the Bootstrapping method is to determine the significance of some field obtained through compositing (see Chapter 3). For example, say you in think that full moons have a strong impact on geopotential heights over Fort Collins, CO. You check this by calculating the average 500 hPa geopotential height on all full moon days (408 between 1979-2011) and obtain 5689.5m. The average over the entire record is 5696.9m - does this mean that full moons cause the geopotential heights to decrease?

The issue here is that you don't know that whether this difference of -7.4m is significant. That is, could it have just been due to random chance? In this case, our null hypothesis is that full moons have no effect on geopotential heights over Fort Collins, and so the anomaly of -7.4m is just due to random chance.

To test this using a bootstrap approach, we randomly draw a sample of geopotential heights of length 408 from all days between 1979-2011. We then calculate the mean across these 408 days and save it. We repeat this process 50,000 times. After we are done, we have 50,000 averages of 408 days - all under the null hypothesis. An example is shown in **Fig. 2.17**, where the gray line denotes the distribution of the 50,000 averages. Comparing our calculated geopotential heights under full moons, we see that they fall well within the bootrap samples. This tells us that we shouldn't reject the null hypothesis, since our actual results are well within the possibility of random chance. If our calculated value fell outside of the 95% bounds of the bootstrap samples, we would instead reject the null hypothesis and investigate further.

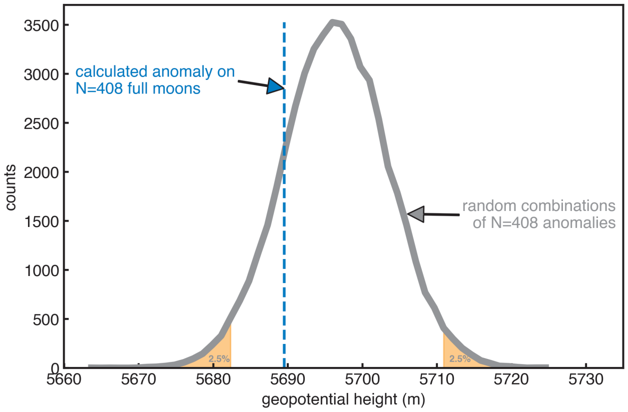

Figure 2.17 Distribution of 50,000 random averages of Fort Collins geopotential heights ($N = 408$).

2.11.3 Resampling via Jackknife

The jackknife method predates the bootstrap method and is a linear approximation of the bootstrap. It is a way of getting uncertainty estimates (or measuring the variance) of a particular statistic of your sample. The way it works is quite simple. You systematically remove one value from your sample (i.e. the ith value), calculate the statistic of interest (call it s_i), then put the value back into the sample and remove the next value, calculate the statistic of interest...and on and on. In the end, you are left with many estimates of your statistic of interest, and from these, you can estimate its variance in the following way.

$$\mathbf{Var}(s) = \frac{N-1}{N} \sum_{i=1}^{N} \left(s_i - \overline{s}_{(.)}\right)^2 \tag{2.126}$$

where $\overline{s}_i$ is the estimate of statistic s leaving out the ith value and $\overline{s}_{(.)}$ is the average estimate of s over all leave-one-out estimates as given by

$$\overline{s}_{(.)} = \frac{1}{N} \sum_{i}^{N} s_i \tag{2.127}$$